Retrieving exoplanet atmospheric parameters using random forest regression

Patcharawee Munsaket[1*], **Supachai Awiphan**[2], **Poemwai Chainakun**[1,3] **and Eamonn Kerins**[4]

[1]School of Physics, Institute of Science, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand
[2]National Astronomical Research Institute of Thailand (Public Organization), 260 Moo 4, Donkaew, Mae Rim, Chiang Mai, 50180, Thailand
[3]Centre of Excellence in High Energy Physics and Astrophysics, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand
[4]Jodrell Bank Centre for Astrophysics, University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom

*E-mail: m6200343@g.sut.ac.th

Abstract. Understanding of exoplanet atmospheres can be extracted from the transmission spectra using an important tool based on a retrieval technique. However, the traditional retrieval method (e.g. MCMC and nested sampling) consumes a lot of computational time. Therefore, this work aims to apply the random forest regression, one of the supervised machine learning technique, to retrieve exoplanet atmospheric parameters from the transmission spectra observed in the optical wavelength. We discovered that the random forest regressor had the best accuracy in predicting planetary radius ($R^2_{Fit} = 0.999$) as well as acceptable accuracy in predicting planetary mass, temperature, and metallicity of planetary atmosphere. Our results suggested that the random forest regression consumes significantly less computing time while gives the predicted results equivalent to those of the nested sampling PLATON retrieval.

1. Introduction

Characterization of planetary atmospheres is a rapidly developing area in the exoplanet field. Transmission spectroscopy is the method commonly used to study exoplanetary atmospheres by measuring the variation of transit depths with wavelengths that depends on properties of elements in the planetary atmosphere [1]. The characteristics of the exoplanet's atmosphere can be obtained from the atmospheric retrieval process. However, the traditional retrieval processes (e.g. MCMC [2] and nested sampling [3]) is a time-consuming procedure. In order to reduce the computational time, machine learning techniques have been implemented for retrieving exoplanet atmosphere parameters. For example, random forest regression, a supervised machine learning technique, has previously been used to approximate the variables of an exoplanet's atmosphere using Hubble Space Telescope Wild Field Camera 3 data (WFC3) [4].

In this paper, random forest regression is employed to retrieve atmospheric properties from simulated photometric observational data of hot Jupiters. The PLanetary Atmospheric Transmission for Observer Noobs (`PLATON`) is used to generate exoplanetary transmission spectra [5]. The spectra then are weighted with the filter transmissions. The atmospheric parameters are estimated using random forest regression model in python `scikit-learn` package [6]. The results are compared with those obtained from nested sampling retrieval in `PLATON`. The workflow is illustrated in figure 1.

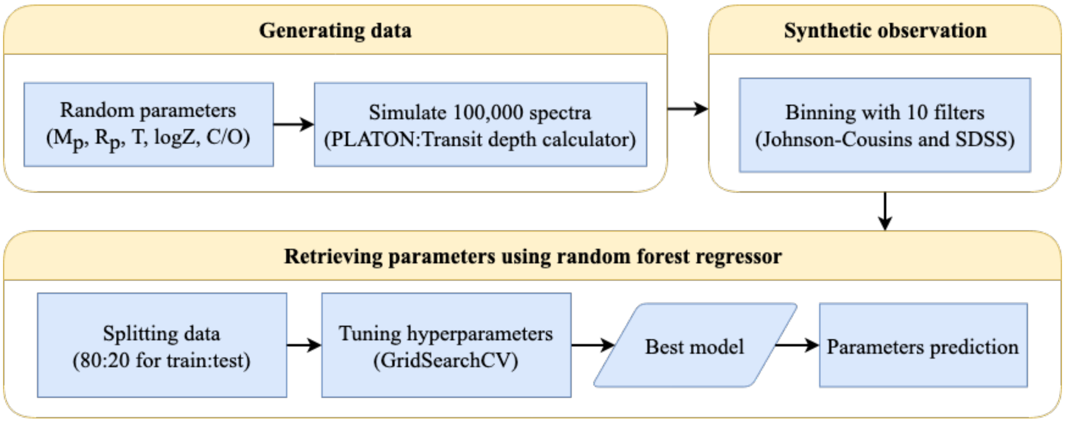

Figure 1. The flowchart for retrieving exoplanet atmospheric parameters from the synthetic transmission spectra in the optical wavelength using a random forest regressor.

2. Synthesize transmission spectra

In this work, only hot Jupiters, gas giant exoplanets with very short orbital periods (P < 10 days), are focused [7]. We investigated five atmospheric parameters including planet mass (M_p), planet radius (R_p), planet atmosphere temperature (T), metallicity of planetary atmosphere ($\log Z$), and atmospheric carbon to oxygen ratio (C/O). `PLATON` forward model is used to generate 100,000 hot Jupiter transmission spectra, by uniformly, randomly selecting the values of these parameters within the range shown in table 1. The adopted planetary mass and radius values are within 68 and 95 confident intervals of the mass-radius relationship of the Jovian exoplanets, respectively [8]. The planets are assumed to orbit a solar-analog star.

Table 1. The input parameters for generating the spectra using the transit depth calculator module in `PLATON`. Note that M_p is in Jupiter mass (M_J) and R_p is in earth radius ($R_\oplus$).

Parameters		Value	References
Planetary mass	(M_p)	$0.3M_J - 3M_J$	[9, 8]
Planetary radius	(R_p)	$8R_\oplus - 20R_\oplus$	[10, 8]
Planetary atmospheric temperature	(T)	1300 – 3000 K	[11]
Metallicity of planetary atmosphere	($\log Z$)	-1 – 3	[5]
Atmospheric Carbon-Oxygen ratio	(C/O)	0.35 – 1	[5]

To date, a number of transit observations are performed from the ground using broad-band optical filters. In this work, The five Johnson-Cousins (U, B, V, R, I) and five Sloan (u', g', r', i', z') optical filters are selected. To synthesize transmission spectra of optical observations, the transmission curve of the filters are used to weight the transmission spectra simulated from `PLATON`, as shown in figure 2. The Johnson-Cousins transmission curves are obtained from photometric transmission curves class in PyAstronomy package [12]. The Sloan filter profile is obtained from Sloan Digital Sky Survey Data Release 7 (SDSS Data Release 7: http://classic.sdss.org/dr7/instruments/imager/#filters). The weighted transmission curves are used as the transit depths that could be obtained from the corresponding filters.

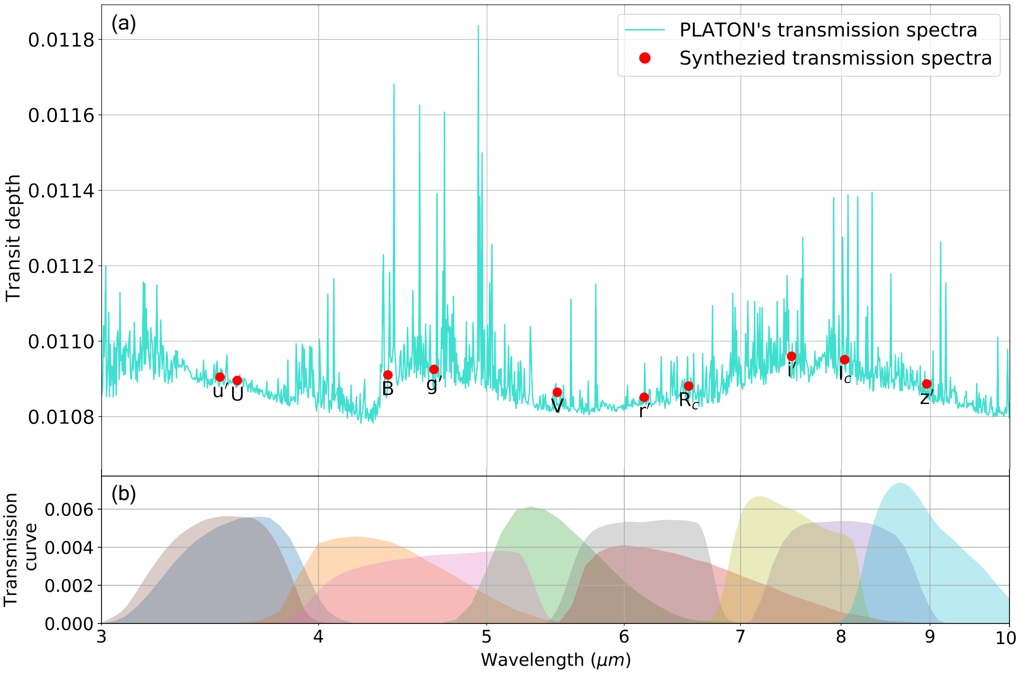

Figure 2. A synthesized transmission spectrum (red) obtained by simulated transmission spectra from PLATON (turquoise) weighted by the transmission curve of the filters (b).

3. Random forest regressor

In this work, the random forest regression is used to retrieve transmission spectra. From 100,000 synthetic transmission spectra of 10 Johnson-Cousins and Sloan filters, the spectra were split into two groups: 80,000 for training and 20,000 for testing. Ten binned transit depths are the model features while five planetary parameters (M_p, R_p, T, $\log Z$ and C/O) are the labels. The hyper-parameters tuning is performed using the GridSearchCV package in scikit-learn to optimize the appropriate values of the number of trees (number of estimators) and the maximum depth of each tree [6]. The best cross validation score is $R^2 = 0.5949$ for the maximum depth of 320 and the number of estimators of 2,560 (figure 3), hence these values are used in our regressor model to predict the planetary parameters from 20,000 spectra in our test data set.

The comparisons between real and guess values of five atmospheric variables are shown in light blue points in figure 4. Their coefficients of determination (R^2_{Fit}) are reported in table 2. The model has the best accuracy in predicting R_p with R^2_{Fit} of 0.999 and moderate accuracy in predicting M_p, T, and $\log Z$ with R^2_{Fit} of 0.742, 0.571 and 0.638, respectively. However, the model cannot predict the C/O ratio as the transmission features induced by the C/O variation are not significant enough in the optical bands.

4. Comparing results with the PLATON retrieval model

The results from our random forest regressor model is compared to those obtained from the nested sampling fitting model in the PLATON. A sample of 100 spectra is simulated and retrieved with the `PLATON`. The number of live points in nested sampling model are selected to be 50 and 100. In order to compare the computaional time, all models perform on a machine with a Core i9-10900 CPU and 32GB of RAM that runs Ubuntu 18.04.2 LTS. The results are shown in figure 4 and table 2. The random forest regressor performs 160,000 and 270,000 times faster than the `PLATON` nested sampling model with 50 and 100 live points, respectively. Judging by R^2_{Fit}, the accuracy obtained from the regressor model are comparable to that of the `PLATON` retrievals.

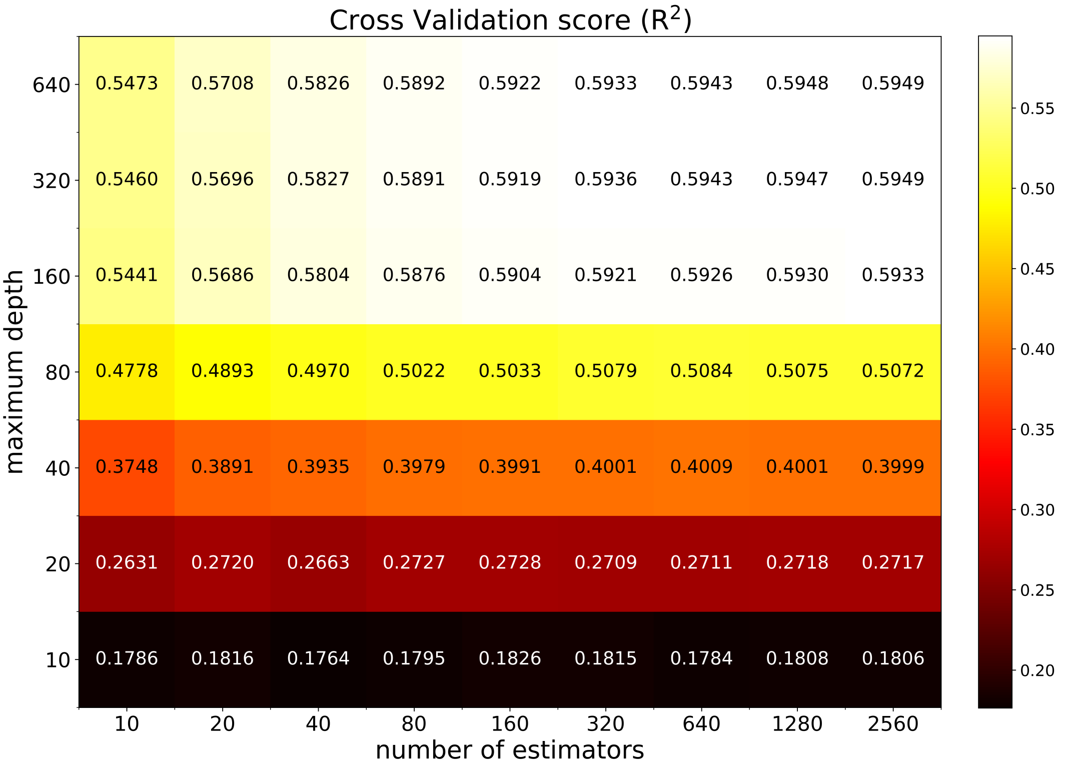

Figure 3. Cross validation score (R^2) obtained from hyper-parameter tuning using `GridSearchCV` package.

Table 2. Performance of our random forest regressor model compared to that of the `PLATON` retrievals with 50 and 100 live points. Note that the models are run on a machine with a Core i9-10900 CPU and 32GB of RAM that runs Ubuntu 18.04.2 LTS.

Approach	**Data sets**	**Time spent**	**Coefficient of determination** (R^2_{Fit})				
		(Hr:Min:Sec)	M_p	R_p	T	$\log Z$	C/O
Random forest regressor	20,000	0:0:14	0.742	0.999	0.571	0.638	0.129
PLATON (nlive=50)	100	3:12:8	0.780	0.999	0.546	0.659	-0.842
PLATON (nlive=100)	100	5:16:45	0.791	0.999	0.616	0.591	-0.604

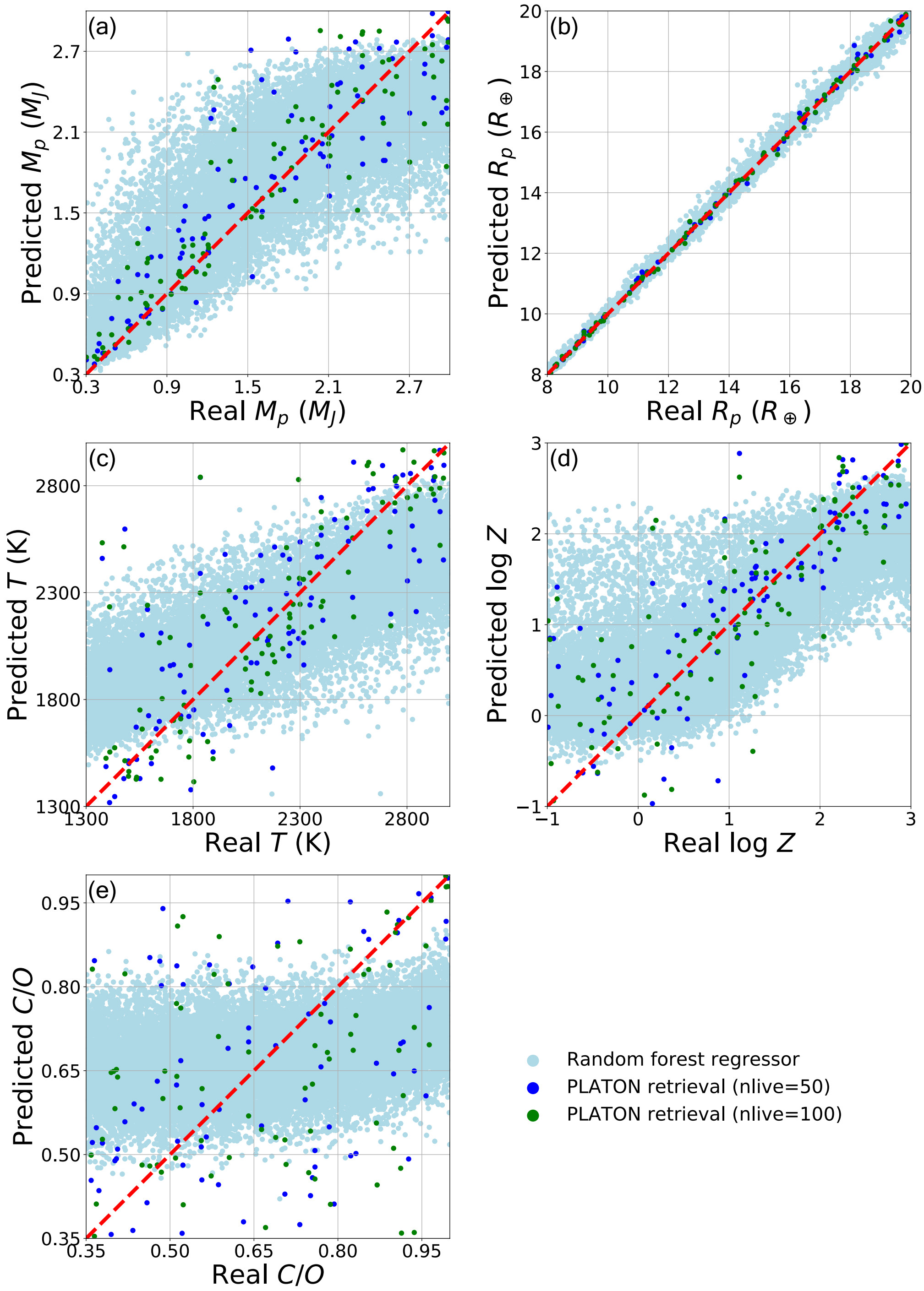

Figure 4. Real and predicted values of five planetary parameters: (a) planetary mass, (b) planetary radius, (c) planetary atmospheric temperature, (d) metalicity of planetary atmosphere and (e) Carbon to Oxygen ratio in the atmosphere, of 20,000 test set using random forest regressor (lightblue), the `PLATON` nested sampling retrival with number of live points of 50 (blue) and 100 (green). The red dashed lines show prefect prediction lines.

5. Conclusion

Transmission spectra of hot-Jupiter atmosphere are simulated using `PLATON` package. The spectra are weighted with 10 Johnson-Cousin and Sloan filter profiles. The random forest regressor are used to estimate five planetary parameters: M_p, R_p, T, $\log Z$ and C/O. The model provides the best cross validation score when the number of estimators is 2560 and the maximum depth is 320. The regressor can precisely predict the planetary radius with $R^2_{Fit} = 0.999$ and has intermediate accuracy in predicting M_p, T, and $\log Z$ values because of the planetary radius can be retrieved using the baseline of transit depths, while the other characteristics can be retrieved from the transit depth features in the specific filters. However, the model cannot predict the C/O correctly because different values of this parameter produce very small variation in optical spectra. We conclude that the random forest regressor can be used to estimate planetary parameters from optical transmission spectra with the prediction accuracy approximately as same as that of the traditional retrieval method, but with more than 100,000 times faster.

Acknowledgements

This work was supported by the Development and Promotion of Science and Technology Talents Project (DPST) - Royal Thai Government Scholarship, Suranaree University of Technology (SUT), and National Astronomical Research Institute of Thailand (NARIT). This research was supported by the Program Management Unit for Human Resources & Institutional Development, Research and Innovation, NXPO [grant number B16F630069].

References

[1] Seager S and Sasselov D 2000 *Astrophys. J.* **537** 916
[2] Foreman-Mackey D, Hogg D W, Lang D and Goodman J 2013 *Publ. Astron. Soc. Pac.* **125** 306
[3] Shaw J, Bridges M and Hobson M 2007 *Mon. Not. R. Astron. Soc.* **378** 1365–70
[4] Márquez-Neila P, Fisher C, Sznitman R and Heng K 2018 *Nat. Astron.* **2** 719–24
[5] Zhang M, Chachan Y, Kempton E M R, Knutson H A and Chang W H 2020 *Astrophys. J.* **899** 27
[6] Pedregosa F *et al.* 2011 *J. Mach. Learn. Res.* **12** 2825–30
[7] Gaudi B S, Seager S and Mallen-Ornelas G 2005 *Astrophys. J.* **623** 472
[8] Chen J and Kipping D 2016 *Astrophys. J.* **834** 17
[9] Stevens D J and Gaudi B S 2013 *Publ. Astron. Soc. Pac.* **125** 933
[10] Borucki W J *et al.* 2011 *Astrophys. J.* **736** 19
[11] Madhusudhan N, Knutson H, Fortney J and Barman T 2014 *Exoplanetary atmospheres* (Arizona: University of Arizona Press) p 739
[12] Czesla S, Schröter S, Schneider C P, Huber K F, Pfeifer F, Andreasen D T and Zechmeister M 2019 Pya: Python astronomy-related packages (*Preprint* `1906.010`)